

Introduction to Bayesian inference for astronomy, 2

Tom Loredo

Cornell Center for Astrophysics and Planetary Science,
Carl Sagan Institute,
& Dept. of Statistics and Data Science, Cornell U.
<http://hosting.astro.cornell.edu/~loredo/>

CASt Summer School — 5–9 June 2023

Recap: Frequentist fundamental principle

*“The most fundamental principle of the statistical paradigm,
its starting point,
is that variation may be described by probability.”*

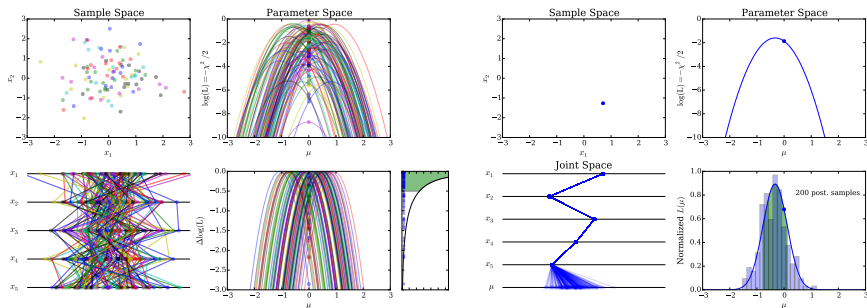
Recap: Bayesian fundamental principle

*The most fundamental principle of the statistical paradigm,
its starting point,
is that **uncertainty** may be described by probability.*

*An important corollary is that, in some settings
—most notably, for **IID replications**,
and for **exchangeable sequences**—
expected variation and individual-case probability
are intimately linked.*

Recap: Confidence vs. credible regions

Hypothetical data vs. hypothetical hypotheses



Find lower/upper functions of the data that make these = 0.683:

$$\text{Cover}(\mu) = \int d^N \vec{x} \, p(\vec{x}|\mu) \, \mathbb{I}[l(\vec{x}) < \mu < u(\vec{x})]$$

$$\text{CredLev}(\vec{x}_{\text{obs}}) = \int d\mu \, p(\mu|\vec{x}_{\text{obs}}) \, \mathbb{I}[L(\vec{x}_{\text{obs}}) < \mu < U(\vec{x}_{\text{obs}})]$$

Bayesian and Frequentist approaches differ

Brad Efron, ASA President (2005)

The 250-year debate between Bayesians and frequentists is unusual among philosophical arguments in actually having *important practical consequences*... The physicists I talked with were really bothered by our 250 year old Bayesian-frequentist argument. Basically there's only one way of doing physics but there seems to be at least two ways to do statistics, and *they don't always give the same answers*...

Broadly speaking, Bayesian statistics dominated 19th Century statistical practice while the 20th Century was more frequentist. What's going to happen in the 21st Century?... I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning...

Roderick Little, ASA President's Address (2005)

Pragmatists might argue that good statisticians can get sensible answers under Bayes or frequentist paradigms; indeed maybe two philosophies are better than one, since they provide more tools for the statistician's toolkit. . . . I am discomforted by this “inferential schizophrenia.” Since the Bayesian (B) and frequentist (F) philosophies *can differ even on simple problems*, at some point decisions seem needed as to which is right. I believe our credibility as statisticians is undermined when we cannot agree on the fundamentals of our subject. . . .

An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that *calibrated Bayes*. . . captures the strengths of both approaches and provides a roadmap for future advances.

[Calibrated Bayes = Bayesian inference within a specified space of models + frequentist-based model checking; Andrew Gelman et al. use *Bayesian data analysis* similarly]

(see TL's arXiv:1208.3035 for discussion/references)

Agenda

① Probability theory for data analysis: Two theorems

② Inference with parametric models

- Parameter Estimation

- Model Uncertainty (Supp.)

③ Quick-looks

- Curve fitting & least squares (2 slides!)

- Bayesian computation menu (1 slide!)

Agenda

① Probability theory for data analysis: Two theorems

② Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

③ Quick-looks

Curve fitting & least squares (2 slides!)

Bayesian computation menu (1 slide!)

The Bayesian recipe

Assess hypotheses by calculating their probabilities $p(H_i | \dots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$\mathcal{C} \equiv$ context, initial set of premises

$$\text{'OR' (sum rule): } P(H_1 \vee H_2 | \mathcal{C}) = P(H_1 | \mathcal{C}) + P(H_2 | \mathcal{C}) - P(H_1, H_2 | \mathcal{C})$$

$$\begin{aligned} \text{'AND' (product rule): } P(H_i, D_{\text{obs}} | \mathcal{C}) &= P(H_i | \mathcal{C}) P(D_{\text{obs}} | H_i, \mathcal{C}) \\ &= P(D_{\text{obs}} | \mathcal{C}) P(H_i | D_{\text{obs}}, \mathcal{C}) \end{aligned}$$

$$\text{'NOT': } P(\overline{H_i} | \mathcal{C}) = 1 - P(H_i | \mathcal{C})$$

Two Important Theorems

Bayes's Theorem (BT)

Consider the *joint probability* for a hypothesis and the observed data, $P(H_i, D_{\text{obs}}|\mathcal{C})$, using the product rule:

$$\begin{aligned}P(H_i, D_{\text{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C}) P(D_{\text{obs}}|H_i, \mathcal{C}) \\&= P(D_{\text{obs}}|\mathcal{C}) P(H_i|D_{\text{obs}}, \mathcal{C})\end{aligned}$$

Solve for the *posterior probability* for H_i (adds a premise!):

$$P(H_i|D_{\text{obs}}, \mathcal{C}) = \frac{P(H_i, D_{\text{obs}}|\mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})} = P(H_i|\mathcal{C}) \frac{P(D_{\text{obs}}|H_i, \mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

(all “for H_i ”)

$$\text{norm. const. } P(D_{\text{obs}}|\mathcal{C}) = \text{prior predictive for } D_{\text{obs}}$$

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ (“suite;” \mathcal{C} asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i | \mathcal{C}) &= \sum_i P(B_i | A, \mathcal{C}) P(A | \mathcal{C}) = P(A | \mathcal{C}) \\ &= \sum_i P(B_i | \mathcal{C}) P(A | B_i, \mathcal{C})\end{aligned}$$

If we do not see how to get $P(A | \mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a “basis”—*extend the conversation*:

$$P(A | \mathcal{C}) = \sum_i P(B_i | \mathcal{C}) P(A | B_i, \mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A | \mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A | \mathcal{C}) = \sum_i P(A, B_i | \mathcal{C})$$

Joseph Blitzstein (Harvard statistician) on LTP (paraphrased):

In most areas of math, when you're stuck, saying, "I wish I knew this or that" doesn't help you. In probability theory, saying "I wish I knew this" suggests what to condition on; then you condition on it, compute *as if* you knew it, and then average over those possibilities.

*I didn't name the law of total probability, but if I had, I would have just called it **wishful thinking**.*

— YouTube lecture on conditional probability (15:48)

LTP example 1: Take \mathcal{C} to specify fair roll of a die, $A =$ “An even number comes up,” $B_i =$ “face i comes up” ($i = 1$ to 6)

$$\begin{aligned}P(A|\mathcal{C}) &= \sum_{i=1}^6 P(A, B_i|\mathcal{C}) \\&= \sum_{i=1}^6 P(B_i|\mathcal{C})P(A|B_i, \mathcal{C}) \\&= \frac{1}{6} \times (0 + 1 + 0 + 1 + 0 + 1) = \frac{1}{2}\end{aligned}$$

LTP example 2: With context \mathcal{C} , take $A = D_{\text{obs}}$, $B_i = H_i$; then

$$\begin{aligned}P(D_{\text{obs}}|\mathcal{C}) &= \sum_i P(D_{\text{obs}}, H_i|\mathcal{C}) \\&= \sum_i P(H_i|\mathcal{C})P(D_{\text{obs}}|H_i, \mathcal{C})\end{aligned}$$

prior predictive for $D_{\text{obs}} =$ Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Tabular/diagrammatic Bayesian inference

Simplest case: *Binary classification*

- 2 hypotheses: $\{C, \overline{C}\}$
- 2 possible data values: $\{-, +\}$

Concrete example: You test positive (+) for a medical condition. Do you have the condition (C) or not (\overline{C})?

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
 - Test is 80% accurate if you have the condition:
 $P(+|C, \mathcal{C}) = 0.8$ (“sensitivity”)
 - Test is 95% accurate if you are healthy:
 $P(-|\overline{C}, \mathcal{C}) = 0.95$ (“specificity,” $1 - p(\text{false } +)$)

Numbers roughly correspond to mammography screening for breast cancer in asymptomatic women

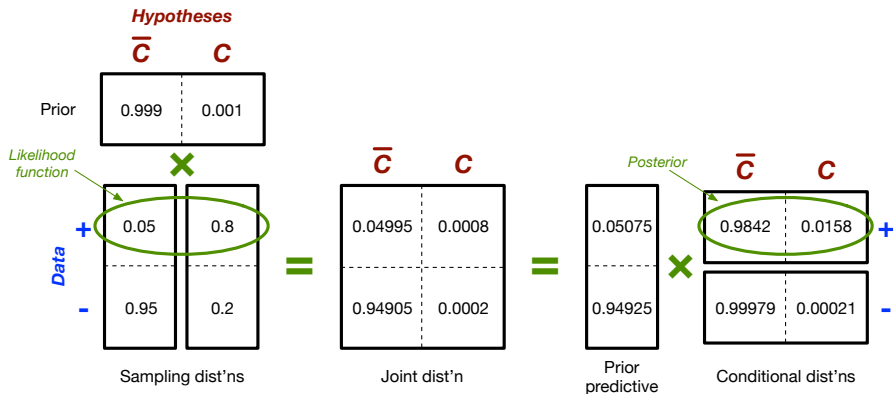
Tabular calculation

Hypothesis H_i	Prior $\pi_i \equiv p(H_i)$	Likelihood $\mathcal{L}_i \equiv p(+ H_i)$	Joint $\pi_i \times \mathcal{L}_i$	Posterior $p(H_i +)$
\overline{C}	0.999	0.05	0.04995	0.9842
C	0.001	0.8	0.0008	0.0158
Sums:	1.0	NA	0.05075 $= p(+)$	1.0

Inference as manipulation of the joint distribution

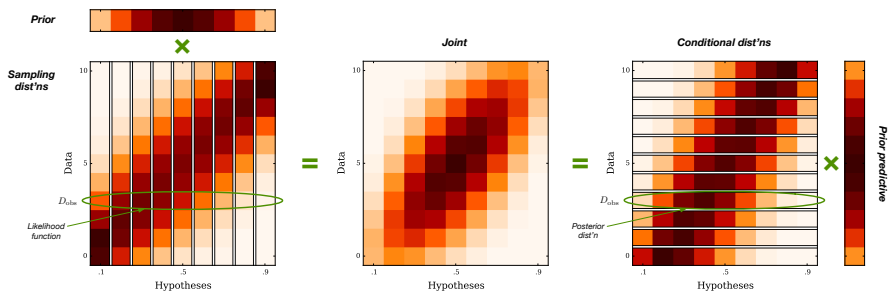
Bayes's theorem in terms of the *joint distribution*:

$$P(H_i|\mathcal{C}) \times P(D_{\text{obs}}|H_i, \mathcal{C}) = P(H_i, D_{\text{obs}}|\mathcal{C}) = P(H_i|D_{\text{obs}}, \mathcal{C}) \times P(D_{\text{obs}}|\mathcal{C})$$



Larger discrete case: Flip a coin 10 times (Binomial inference)

- 9 hypotheses: Prob. for heads is $\alpha = 0.1, 0.2, \dots, 0.9$
- 11 possible data values: Number of heads, $n = 0, 1, \dots, 10$
- Adopt a prior concentrated around $\alpha = 0.5$, with some spread



Joint ranks “possible worlds”— (H_i, D) pairs—before observing data.

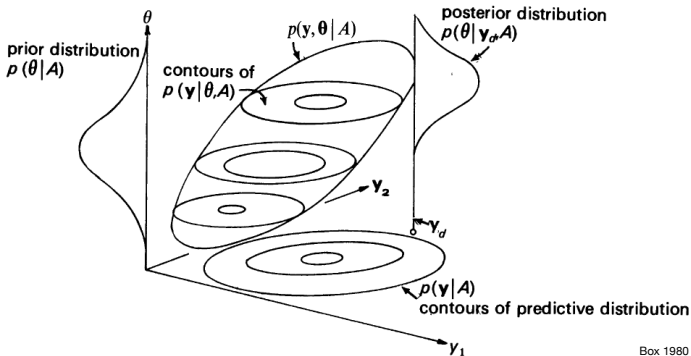
Posterior adjusts the ranks once we learn the relevant possibilities must have $D = D_{obs}$.

Continuous data, parameter spaces

Prior \times sampling distribution gives the *joint dist'n*:

$$p(\theta, D) = p(\theta) \times p(D|\theta)$$

Conditioning on $D = D_{\text{obs}}$ gives the posterior:



Box 1980

Components of Bayes's theorem for a problem with a 1-D parameter space (θ) and a 2-D sample space (\mathbf{y}), with observed data \mathbf{y}_d , and modeling assumptions A

Recap of key ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis \propto ability of hypothesis to *predict* the observed data

Law of total probability

$$p(\text{Hypotheseses} \mid \text{Data}) = \sum p(\text{Hypothesisis} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Agenda

① Probability theory for data analysis: Two theorems

② Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

③ Quick-looks

Curve fitting & least squares (2 slides!)

Bayesian computation menu (1 slide!)

Inference with parametric models

Models M_i ($i = 1$ to N), each with parameters θ_i , each imply a *sampling dist'n* (conditional predictive dist'n for possible data):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the *observed* data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty).

Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript: $D = D_{\text{obs}}$.

Classes of problems

Single-model inference

Premise = choice of single model (specific i)

Parameter estimation: What can we say about θ_i or $f(\theta_i)$?

Prediction: What can we say about future data D' ?

Multi-model inference

Premise = $\{M_i\}$

Model comparison/choice: What can we say about i ?

Model averaging:

- *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D' , accounting for model uncertainty?

Model checking

Premise = $M_1 \vee$ “all” alternatives

Is M_1 adequate? (predictive tests, calibration, robustness)

Parameter estimation

Problem statement

\mathcal{C} = Model M with parameters θ (+ any add'l info)

H_i = statements about θ ; e.g. " $\theta \in [2.5, 3.5]$," or " $\theta > 0$ "

Probability for any such statement can be found using a *probability density function* (PDF) for θ :

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \dots) &= f(\theta) d\theta \\ &= p(\theta | \dots) d\theta \end{aligned}$$

Posterior probability density

$$p(\theta | D, M) = \frac{p(\theta | M) \mathcal{L}(\theta)}{\int d\theta p(\theta | M) \mathcal{L}(\theta)}$$

Summaries of posterior

- “Best fit” values:
 - ▶ *Mode*, $\hat{\theta}$, maximizes $p(\theta|D, M)$
 - ▶ *Posterior mean*, $\langle \theta \rangle = \int d\theta \theta p(\theta|D, M)$
- Uncertainties:
 - ▶ *Credible region* Δ of probability C :
 $C = P(\theta \in \Delta|D, M) = \int_{\Delta} d\theta p(\theta|D, M)$
Highest Posterior Density (HPD) region has $p(\theta|D, M)$ higher inside than outside
 - ▶ Posterior standard deviation, variance, covariances
- Marginal distributions
 - ▶ Interesting parameters ϕ , nuisance parameters η
 - ▶ *Marginal dist'n* for ϕ : $p(\phi|D, M) = \int d\eta p(\phi, \eta|D, M)$

Estimating a normal mean

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is known $\rightarrow I = (\sigma, M)$.

Parameter space: μ ; seek $p(\mu|D, \sigma, M)$

Likelihood

$$\begin{aligned}\mathcal{L}(\mu) &\equiv p(D|\mu, \sigma, M) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(d_i - \mu)^2 / 2\sigma^2}; \quad \sigma = 1 \\ &\propto \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right)\end{aligned}$$

Likelihood function is a Gaussian function at \bar{d} , width $w = \sigma/\sqrt{N}$

Informative conjugate prior

Use a normal prior, $\mu \sim \mathcal{N}(\mu_0, w_0^2)$

Conjugate because the posterior turns out also to be normal

Posterior

Normal $\mathcal{N}(\tilde{\mu}, \tilde{w}^2)$, but mean shifts towards prior, std. deviation decreases (reflecting add'l info from the prior)

Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large; then

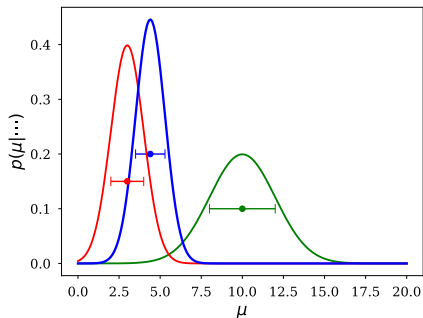
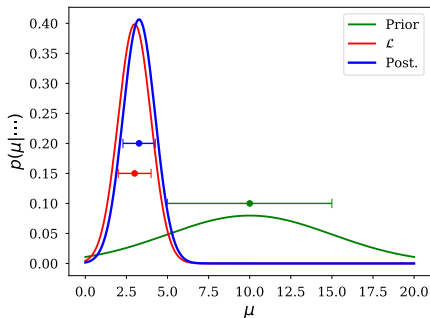
$$\begin{aligned}\tilde{\mu} &= \bar{d} + B \cdot (\mu_0 - \bar{d}) \\ \tilde{w} &= w \cdot \sqrt{1 - B}\end{aligned}$$

Principle of stable estimation/precise measurement — “If observations are precise. . . relative to the prior, then the form and properties of the prior distribution have negligible influence on the posterior distribution.”

Edwards, Lindman, and Savage (1963), ‘Bayesian Statistical Inference for Psychological Research,’ reprinted in *Breakthroughs in Statistics*

Conjugate normal examples:

- Data have $\bar{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$



Supplement:

- Binomial example
 - ▶ Bernoulli trials: Bernoulli process & binomial sampling dist'ns
 - ▶ Beta-binomial conjugate model
- Normal example, cont'd
 - ▶ Analytical details for normal example
 - ▶ Sufficiency; sample mean and variance as sufficient statistics
 - ▶ Handling σ uncertainty by marginalizing over σ ; Student's t distribution

Nuisance parameters and marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b .

We have additional data just about b .

What do the data tell us about s ?

Marginal posterior distribution

To summarize implications for s , accounting for b uncertainty, *marginalize*:

$$\begin{aligned} p(s|D, M) &= \int db \, p(s, b|D, M) \\ &\propto p(s|M) \int db \, p(b|s, M) \mathcal{L}(s, b) \\ &= p(s|M) \mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function* for s :

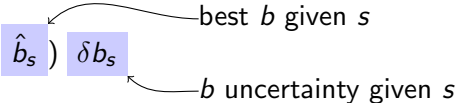
$$\mathcal{L}_m(s) \equiv \int db \, p(b|s) \mathcal{L}(s, b)$$

Maximum likelihood suggests instead computing the *profile likelihood*:

$$\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s), \quad \hat{b}_s = \text{best } b \text{ given } s$$

Marginalization vs. profiling

For insight: Suppose the prior is broad compared to the likelihood
→ for a fixed s , we can accurately estimate b with max likelihood \hat{b}_s , with small uncertainty δb_s .

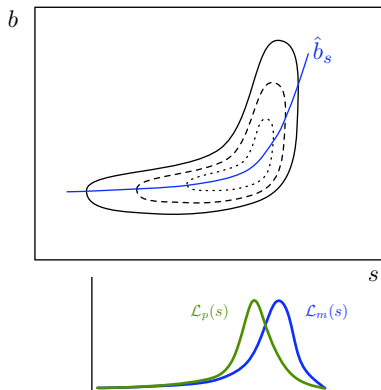
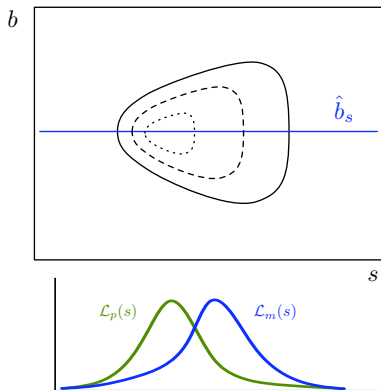
$$\begin{aligned}\mathcal{L}_m(s) &\equiv \int db \, p(b|s) \mathcal{L}(s, b) \\ &\approx p(\hat{b}_s|s) \mathcal{L}(s, \hat{b}_s) \delta b_s\end{aligned}$$


Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s, \hat{b}_s)$ gets weighted by a *parameter space volume factor*

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}$, $\sigma_s^2 = \sigma_r^2 + \sigma_b^2$

Background *subtraction* is a special case of background *marginalization*.

Flared/skewed/bannana-shaped: \mathcal{L}_m and \mathcal{L}_p differ



General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$. Otherwise, they will likely *differ*.

In *measurement error problems* the difference can be dramatic

The on/off problem for Poisson counting data

Basic problem

- Look off-source; unknown background rate b
Count N_{off} photons in interval T_{off}
- Look on-source; rate is $r = s + b$ with unknown signal s
Count N_{on} photons in interval T_{on}
- Infer s

Conventional solution

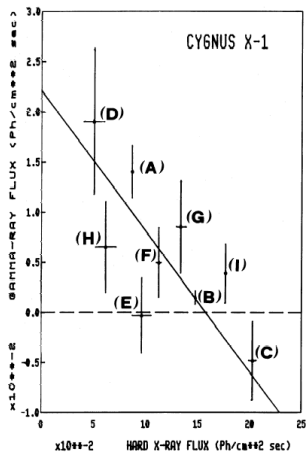
$$\begin{aligned}\hat{b} &= N_{\text{off}} / T_{\text{off}}; & \sigma_b &= \sqrt{N_{\text{off}}} / T_{\text{off}} \\ \hat{r} &= N_{\text{on}} / T_{\text{on}}; & \sigma_r &= \sqrt{N_{\text{on}}} / T_{\text{on}} \\ \hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2}\end{aligned}$$

But \hat{s} can be *negative*!

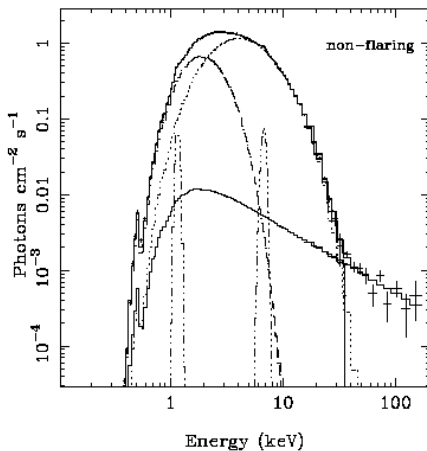
Examples

Spectra of X-ray sources

Bassani et al. 1989

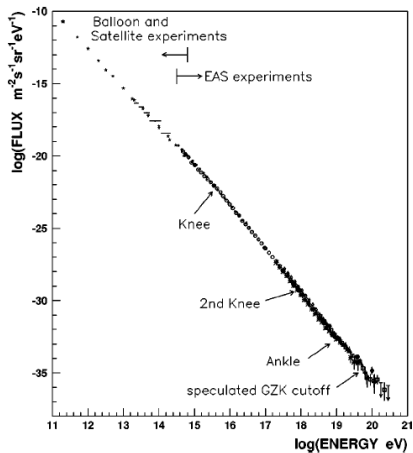


Di Salvo et al. 2001

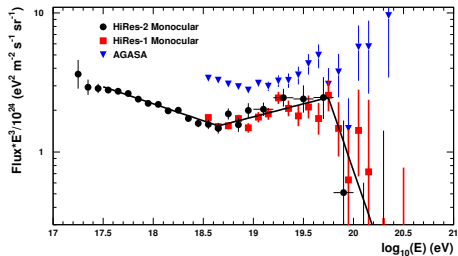


Spectrum of ultrahigh-energy cosmic rays

Nagano & Watson 2000



HiRes Team 2007



N is never large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

— Andrew Gelman (blog entry, 31 July 2005)

N is never large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

Similarly, you never have quite enough money. But that's another story.

— Andrew Gelman (blog entry, 31 July 2005)

Bayesian solution to on/off problem

The likelihood function is a product of separate Poisson distributions for the off-source and on-source data:

$$\mathcal{L}(s, b) = \frac{(bT_{\text{off}})^{N_{\text{off}}}}{N_{\text{off}}!} e^{-bT_{\text{off}}} \times \frac{[(s+b)T_{\text{on}}]^{N_{\text{on}}}}{N_{\text{on}}!} e^{-(s+b)T_{\text{on}}}$$

Adopting flat priors for (s, b) , the joint posterior is

$$p(s, b | N_{\text{on}}, N_{\text{off}}, \mathcal{C}) \propto (s+b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-sT_{\text{on}}} e^{-b(T_{\text{on}}+T_{\text{off}})}$$

Note if $b = 0$, the (normalized) posterior distribution is a gamma distribution,

$$p(s, b = 0 | N_{\text{on}}, N_{\text{off}}, \mathcal{C}) = \frac{T_{\text{on}}(sT_{\text{on}})^{N_{\text{on}}}}{N_{\text{on}}!} e^{-sT_{\text{on}}}$$

Now marginalize over b ;

$$\begin{aligned} p(s|N_{\text{on}}, N_{\text{off}}, \mathcal{C}) &= \int db \, p(s, b | N_{\text{on}}, \mathcal{C}) \\ &\propto \int db \, (s + b)^{N_{\text{on}}} b^{N_{\text{off}}} e^{-sT_{\text{on}}} e^{-b(T_{\text{on}} + T_{\text{off}})} \end{aligned}$$

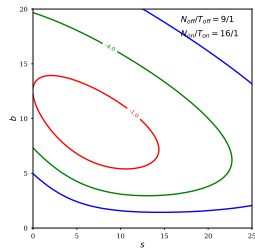
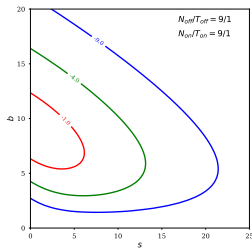
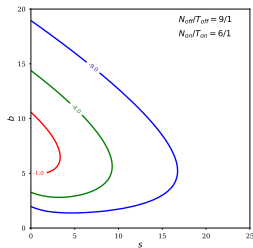
Expand $(s + b)^{N_{\text{on}}}$ and do the resulting Γ integrals:

$$\begin{aligned} p(s|N_{\text{on}}, N_{\text{off}}, \mathcal{C}) &= \sum_{i=0}^{N_{\text{on}}} C_i \frac{T_{\text{on}} (sT_{\text{on}})^i e^{-sT_{\text{on}}}}{i!} \\ C_i &\propto \left(1 + \frac{T_{\text{off}}}{T_{\text{on}}}\right)^i \frac{(N_{\text{on}} + N_{\text{off}} - i)!}{(N_{\text{on}} - i)!} \end{aligned}$$

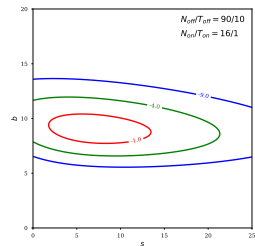
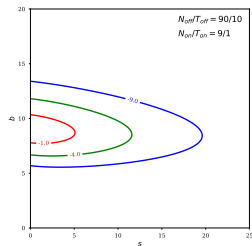
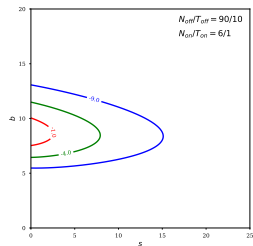
Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)

Example on/off joint PDFs

$$T_{\text{on}} = T_{\text{off}} = 1$$

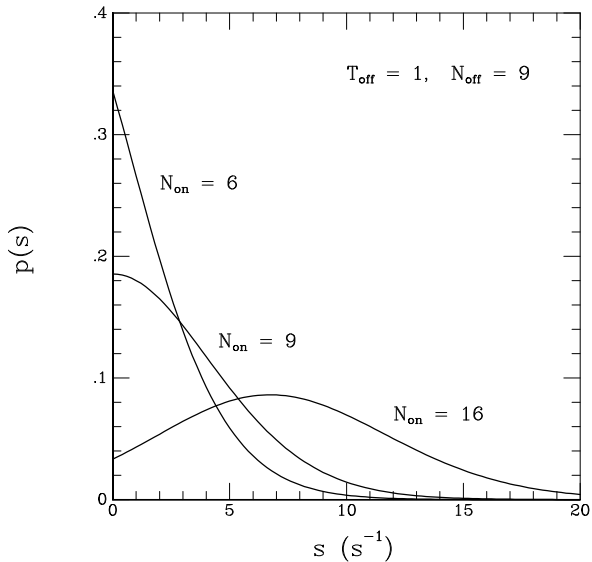


$$T_{\text{on}} = 1, T_{\text{off}} = 10$$



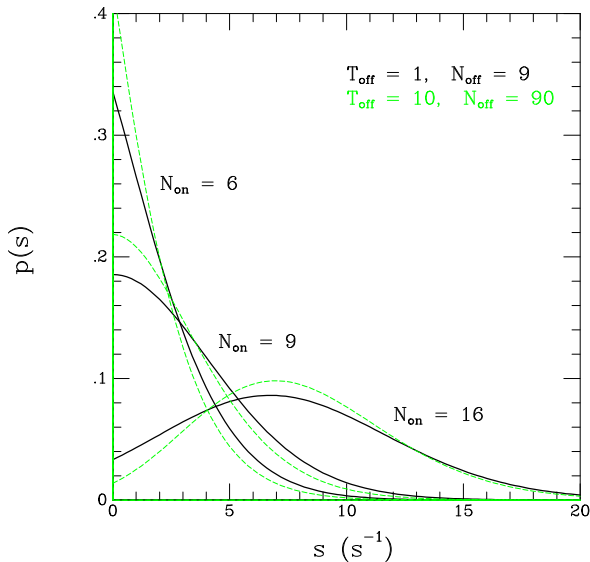
Example on/off marginal PDFs—Short integrations

$$T_{\text{on}} = T_{\text{off}} = 1$$



Example on/off marginal PDFs—Long background integrations

$$T_{\text{on}} = 1, T_{\text{off}} = 10$$



Supplement:

- Analytical details for Poisson dist'n inference
- Gamma-Poisson conjugate model
- Alternative (equivalent) solution to the on/off problem
- Multibin case

Many roles for marginalization

Eliminate nuisance parameters

$$p(\phi|D, M) = \int d\eta \, p(\phi, \eta|D, M)$$

Propagate uncertainty

Model has parameters θ ; what can we infer about $F = f(\theta)$?

$$\begin{aligned} p(F|D, M) &= \int d\theta \, p(F, \theta|D, M) = \int d\theta \, p(\theta|D, M) p(F|\theta, M) \\ &= \int d\theta \, p(\theta|D, M) \delta[F - f(\theta)] \quad [\textit{single-valued case}] \end{aligned}$$

Prediction

Given a model with parameters θ and present data D , predict future data D' (e.g., for *experimental design*):

$$p(D'|D, M) = \int d\theta \, p(D', \theta|D, M) = \int d\theta \, p(\theta|D, M) p(D'|\theta, M)$$

Model comparison. . .

Many roles for marginalization

Many composite hypotheses are of interest. . .

Credible regions

$$p(\theta \in \Delta | D, M) = \int_{\Delta} d\theta \, p(\theta | D, M)$$

Eliminate nuisance parameters

$$p(\phi | D, M) = \int d\eta \, p(\phi, \eta | D, M)$$

Propagate uncertainty

Model has parameters θ ; what can we infer about $F = f(\theta)$?

$$\begin{aligned} p(F | D, M) &= \int d\theta \, p(F, \theta | D, M) = \int d\theta \, p(\theta | D, M) p(F | \theta, M) \\ &= \int d\theta \, p(\theta | D, M) \delta[F - f(\theta)] \quad [\textit{single-valued case}] \end{aligned}$$

Prediction

Given a model with parameters θ and present data D , predict future data D' (e.g., for *experimental design*):

$$p(D'|D, M) = \int d\theta \, p(D', \theta|D, M) = \int d\theta \, p(\theta|D, M) p(D'|\theta, M)$$

Hierarchical modeling (*graphical models, multilevel models*)

Learn population parameters by marginalizing over latent object parameters

Learn an object's parameters by marginalizing over pop'n model and other objects' parameters \rightarrow *shrinkage*

Model uncertainty & multi-model inference...

Agenda

① Probability theory for data analysis: Two theorems

② Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

③ Quick-looks

Curve fitting & least squares (2 slides!)

Bayesian computation menu (1 slide!)

Model uncertainty & multi-model inference

Supplement:

- Odds and Bayes factors: Compare models using *marginal* (average) likelihoods, not *maximum* likelihoods
- Bayesian Ockham's razor and Ockham factors
- Bayesian model averaging

In a nutshell:

Marginal likelihood for model M_i :

$$Z_i \equiv p(D|M_i) = \int d\theta_i \, p(\theta_i|M) \mathcal{L}_i(\theta_i)$$

Bayes factor $B_{ij} \equiv Z_i/Z_j$ (ratio of *average*, not *max* likelihoods)

Can write $Z_i = \mathcal{L}_i(\hat{\theta}_i) \cdot \Omega_i$ with *Ockham factor*

$\Omega_i \approx \delta\theta/\Delta\theta = (\text{posterior volume})/(\text{prior volume})$

A misconception

Bayesian data analysis gets its name from **Bayes's theorem**:

$$\begin{aligned} p(\theta|D_{\text{obs}}) &= \frac{p(\theta) p(D_{\text{obs}}|\theta)}{p(D_{\text{obs}})} \\ &= \frac{p(\theta) \mathcal{L}(\theta)}{p(D_{\text{obs}})} \end{aligned}$$

So it's basically about **modulating maximum likelihood with priors**...

Bayesian data analysis gets its name from *Bayes's theorem*.

$$\begin{aligned} p(\theta|D_{\text{obs}}) &= \frac{p(\theta) p(D_{\text{obs}}|\theta)}{p(D_{\text{obs}})} \\ &= \frac{p(\theta) \mathcal{L}(\theta)}{p(D_{\text{obs}})} \end{aligned}$$

So it's basically about *modulating maximum likelihood with priors*...

Computing the posterior is just the *starting point*—we then have to do calculations, using all of probability theory to get answers to our questions from the posterior.

We'll find ourselves using the *law of total probability* over and over again—**marginalization** (summing/integrating probabilities).

Integrating over parameter space is the key feature distinguishing Bayesian from frequentist data analysis
(frequentist methods typically **optimize** over parameter space)

On the key role of marginalization

Bayesian statistics uses all of probability theory, not just Bayes's theorem, and not even primarily Bayes's theorem. . . . Perhaps the most important theorem for doing Bayesian calculations is the *law of total probability* (LTP) that relates marginal probabilities to joint and conditional probabilities. . . . Arguably, if this approach to inference is to be named for a theorem, "total probability inference" would be a more appropriate appellation than "Bayesian statistics." It is probably too late to change the name. But it is not too late to change the emphasis.

— Loredó (2013)

The key distinguishing property of a Bayesian approach is marginalization instead of optimization, not the prior, or Bayes rule. . . . Broadly speaking, what makes Bayesian approaches distinctive is a posterior weighted marginalization over parameters. . . . Moreover, basic probability theory indicates that marginalization is desirable.

— Wilson (2020), Wilson & Izmailov (2020)

Roles of the prior

Prior has two roles

- Modulate the likelihood to incorporate relevant prior information
- Convert likelihood from “intensity” to “measure”
→ enable accounting for *size of parameter space*

Physical analogy

$$\text{Heat } Q = \int d\vec{r} c_v(\vec{r}) T(\vec{r})$$

$$\text{Probability } P \propto \int d\theta p(\theta) \mathcal{L}(\theta)$$

Maximum likelihood focuses on the “hottest” parameters.

Bayes focuses on the parameters with the most “heat.”

A high- T region may contain little heat if its c_v is low or if its volume is small.

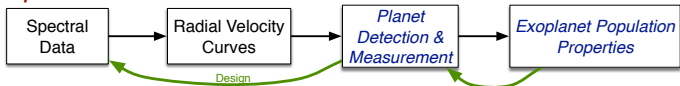
A high- \mathcal{L} region may contain little probability if its prior is low or if its volume is small.

Priors are like initial conditions/boundary conditions in physics: sometimes a nuisance, sometimes crucially important, always required by the theory

By converting likelihood to probability, priors provide two crucial capabilities:

- Accumulation of evidence (learning) → *discovery chains*

Exoplanets



- Automatic accounting for the sizes of parameter/hypothesis spaces: nuisance parameters, uncertainty propagation, prediction, model comparison. . .

If your problem needs particularly careful and thorough implementation of these capabilities, you should consider Bayesian methods

Supplement:

- Assigning priors
- Rule-based “objective” priors: Jeffreys, reference

Also see Stan’s “Prior Choice Recommendations” Wiki

Theme: Parameter space volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Model marginal likelihoods have parameter space volume factors that can penalize models for unnecessary complexity
- Prediction, uncertainty propagation, model averaging. . .

Many virtues of Bayesian methods can be attributed to this accounting for the “size” of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added “by hand”—ignoring Fisher!).

Agenda

① Probability theory for data analysis: Two theorems

② Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

③ Quick-looks

Curve fitting & least squares (2 slides!)

Bayesian computation menu (1 slide!)

Bayesian curve fitting & least squares

Setup

Data $D = \{d_i\}$ are measurements of an underlying function $f(x; \theta)$ at N sample points $\{x_i\}$. Let $f_i(\theta) \equiv f(x_i; \theta)$:

$$d_i = f_i(\theta) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2)$$

We seek to learn θ , or to compare different functional forms (model choice, M)

Likelihood

$$\begin{aligned} p(D|\theta, M) &= \prod_{i=1}^N \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{d_i - f_i(\theta)}{\sigma_i} \right)^2 \right] \\ &\propto \exp \left[-\frac{1}{2} \sum_i \left(\frac{d_i - f_i(\theta)}{\sigma_i} \right)^2 \right] \\ &= \exp \left[-\frac{\chi^2(\theta)}{2} \right] \end{aligned}$$

Posterior

For prior density $\pi(\theta)$,

$$p(\theta|D, M) \propto \pi(\theta) \exp \left[-\frac{\chi^2(\theta)}{2} \right]$$

If you have a least-squares or χ^2 code:

- Treat $\chi^2(\theta)$ as $-2 \log \mathcal{L}(\theta)$
- Bayesian inference amounts to exploration and *numerical integration* (by quadrature or Monte Carlo) of $\pi(\theta)e^{-\chi^2(\theta)/2}$

Forthcoming Python *Parametric Inference Engine* (PIE):

```
class MyData(PredictorSet):
    d1 = SampledGaussianPred(data1, doc="Sampled")
    d2 = BinnedGaussianPred(data2, doc="Binned")

class PowerLaw(SignalModel):
    A = PosParam(1., 'Amplitude')
    alpha = RealParam(range=(-5,-1), 'Index')

    def signal(self, E):
        return self.A * E**self.alpha
```

```
class PowerLawInference(BayesianInference,
                        PowerLaw, MyData):
    def log_prior(self):
        return 0. # const. prior

inf = PowerLawInference()

inf.A.vary()
inf.alpha.step(0., 5., 50)

grid1 = laplace() # Laplace approx.
grid2 = marg()    # Marg. via cubature
```

Agenda

① Probability theory for data analysis: Two theorems

② Inference with parametric models

Parameter Estimation

Model Uncertainty (Supp.)

③ Quick-looks

Curve fitting & least squares (2 slides!)

Bayesian computation menu (1 slide!)

Bayesian computation menu

Large sample size, N : Laplace approximation

- Approximate posterior as multivariate normal $\rightarrow \det(\text{covar})$ factors
- Uses ingredients available in χ^2 /ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$ (better than $O(1/\sqrt{N})$)

Modest-dimensional models ($m \lesssim 10$ to 20)

- Quadrature, cubature, adaptive cubature
- IID Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($m \gtrsim 5$): Non-IID Monte Carlo

- Posterior sampling — create RNG that samples posterior
 - ▶ Markov Chain Monte Carlo (MCMC) is the most general framework
- Nested sampling
- Sequential Monte Carlo (SMC)
- Approximate(ly) Bayesian computation (ABC)/Likelihood-free inference (LFI)
- ...

Recap of key ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis \propto ability of hypothesis to *predict* the observed data

Law of total probability

$$p(\text{Hypotheses} \mid \text{Data}) = \sum p(\text{Hypothesis} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Supplemental Topics

- ① Estimation and model comparison for binary outcomes; probability & frequency
- ② Basic inference with normal errors
- ③ Poisson distribution; the on/off problem
- ④ Model uncertainty
- ⑤ Measurement error & hierarchical/graphical models
- ⑥ Typical sets (from MOO to MOE)
- ⑦ Assigning priors